SALMONN: Towards Generic Hearing Abilities for Large Language Models

Year: 2023 | *Citations: 416* | *Authors: Changli Tang, Wenyi Yu, Guangzhi Sun*

Abstract

Hearing is arguably an essential ability of artificial intelligence (AI) agents in the physical world, which refers to the perception and understanding of general auditory information consisting of at least three types of sounds: speech, audio events, and music. In this paper, we propose SALMONN, a speech audio language music open neural network, built by integrating a pre-trained text-based large language model (LLM) with speech and audio encoders into a single multimodal model. SALMONN enables the LLM to directly process and understand general audio inputs and achieve competitive performances on a number of speech and audio tasks used in training, such as automatic speech recognition and translation, auditory-information-based question answering, emotion recognition, speaker verification, and music and audio captioning etc. SALMONN also has a diverse set of emergent abilities unseen in the training, which includes but is not limited to speech translation to untrained languages, speech-based slot filling, spoken-query-based question answering, audio-based storytelling, and speech audio co-reasoning etc. The presence of cross-modal emergent abilities is studied, and a novel few-shot activation tuning approach is proposed to activate such abilities. To our knowledge, SALMONN is the first model of its type and can be regarded as a step towards AI with generic hearing abilities. The source code, model checkpoints and data are available at https://github.com/bytedance/SALMONN.